Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR

Warren *et al*. 10.1073/pnas.0608512103.

Supporting Information

Files in this Data Supplement:

Supporting Figure 5
Supporting Table 2
Supporting Table 3
Supporting Table 4
Supporting Table 5
Supporting Table 6
Supporting Table 7
Supporting Text

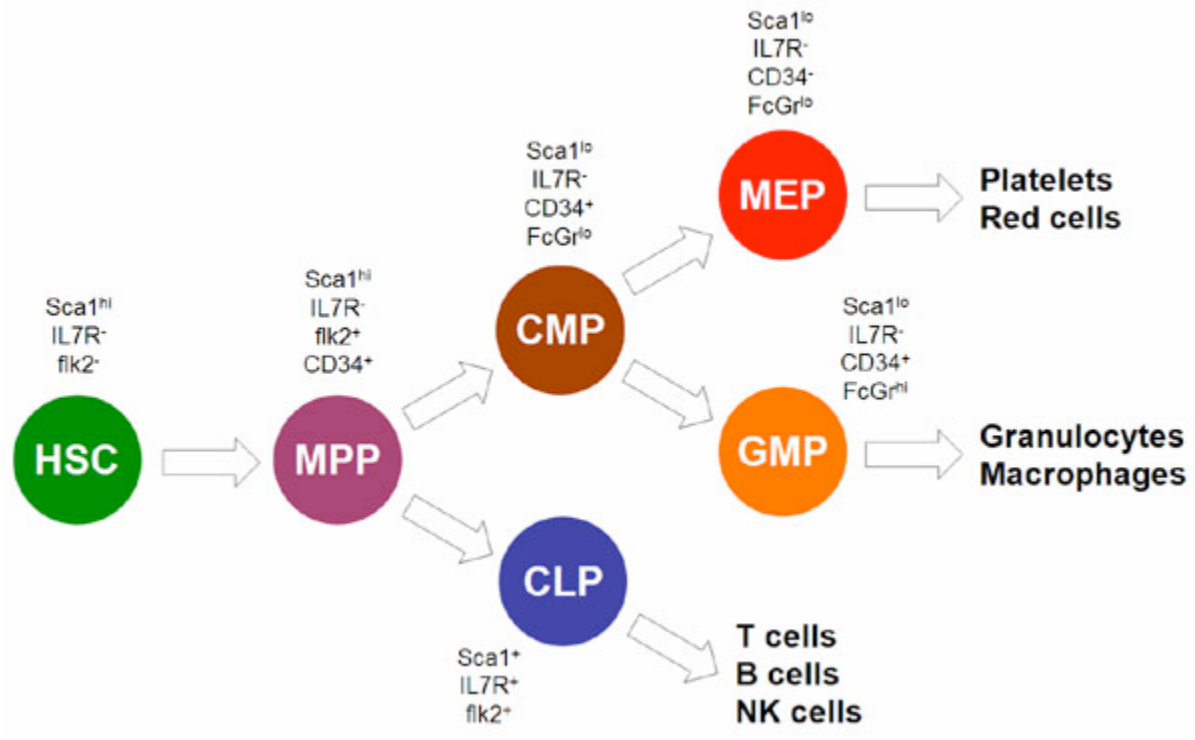

Fig. 5. Early progenitors in the hematopoietic lineage tree, according to the classical model of blood differentiation. Upon activation, self-renewing hematopoietic stem cells (HSCs) give rise to a proliferating population of multipotent progenitors (MPPs). Two more restricted oligopotent precursors are derived from the MPP: the common myeloid progenitor (CMP) and the common lymphoid progenitor (CLP). The CMP population forks into still-more restricted oligopotent precursors: the megakaryocyte-erythroid progenitor (MEP) and the granulocyte-macrophage progenitor (GMP). The CLP develops into the

unipotent precursors of B and T cells directly. Phenotypically, these early progenitors are all positive for the stem cell factor receptor, c-kit, and negative for the lineage-specific markers which characterize more mature cells. The surface marker phenotypes that distinguish the different progenitor types within the Lineage$^-$ c-kit$^+$ compartment are indicated in the figure.

Table 2. False positives

	NTC (n = 6)						No RT (n = 3)		
GAPDH	2	1	1	4	3	1	3	2	2
PU.1	0	0	0	0	0	0	1	0	0

Counts of false-positive wells within NTC and single-cell No RT control panels. The equivalent copies-per-cell readouts would be double these values, as the loaded sample volume is half the RT reaction volume.

Table 3. Descriptive statistics (PU.1)

	HSC	CLP	CMP/flk2$^+$	CMP/flk2$^-$	MEP
Number of samples	21	23	25	24	23
Median cDNAs per cell	6.0	4.1	14.1	3.0	2.0
Mean cDNAs per cell	8.5	5.5	21.7	6.5	3.7
Coefficient of variation	95%	82%	120%	149%	180%
Geometric mean cDNAs per cell	6.0	n/a	14.8	n/a	n/a
Geometric standard deviation	2.3	n/a	2.4	n/a	n/a

Descriptive statistics for the PU.1 expression single-cell data. The geometric mean and geometric standard deviation correspond to the back-transformed mean and standard deviation of the log-transformed data. These two statistics are more informative than the mean and standard deviation for lognormally distributed data. (Their values are mathematically undefined for data sets which include zero values, as was the case with the CLP, CMP/flk2$^-$ and MEP data sets here.)

Table 4. Subset comparisons (PU.1)

	HSC	CLP	CMP/flk2$^+$	CMP/flk2$^-$	MEP
HSC	1.00	0.89	0.00	0.37	0.03
CLP	0.89	1.00	0.00	0.89	0.20
CMP/flk2$^+$	0.00	0.00	1.00	0.00	0.00
CMP/flk2$^-$	0.37	0.89	0.00	1.00	0.67
MEP	0.03	0.20	0.00	0.67	1.00

Results of pairwise Kolmogorov-Smirnov comparison tests between PU.1 expression datasets. The tabulated values are the significance levels assigned by the test to the null hypothesis that the data from the two compared sets come from the same underlying distribution. A low *P* value (<0.05) is evidence that the distributions differ significantly.

Table 5. Descriptive statistics (GAPDH)

	HSC	CLP	CMP/flk2$^+$	CMP/flk2$^-$	MEP
Number of samples	21	23	25	24	23
Median cDNAs per cell	45	26	65	43	45
Mean cDNAs per cell	58	37	72	47	61
Coefficient of variation	86%	102%	58%	63%	78%
Geometric mean cDNAs per cell	43	22	57	39	46
Geometric standard deviation	2.5	3.1	2.3	1.8	2.2

Descriptive statistics for the GAPDH single-cell expression data. The geometric mean and geometric standard deviation correspond to the back-transformed mean and standard deviation of the log-transformed data. These two statistics are more informative than the mean and standard deviation for lognormally distributed data.

Table 6. Subset comparisons (GAPDH)

	HSC	CLP	CMP/flk2$^+$	CMP/flk2$^-$	MEP
HSC	1.00	0.01	0.04	0.93	0.67
CLP	0.01	1.00	0.00	0.07	0.20
CMP/flk2$^+$	0.04	0.00	1.00	0.03	0.06
CMP/flk2$^-$	0.93	0.07	0.03	1.00	0.33
MEP	0.67	0.20	0.06	0.33	1.00

Results of pairwise Kolmogorov-Smirnov comparison tests between GAPDH expression datasets. The tabulated values are the significance levels assigned by the test to the null hypothesis that the data from the two compared sets comes from the same underlying distribution. A low *P* value (<0.05) is evidence that the distributions differ significantly.

Table 7. Normality tests

	Test	HSC	CLP	CMP/flk2$^+$	CMP/flk2$^-$	MEP
Raw data	Shapiro-Wilk	0.00 / 0.00	0.00	0.49 / 0.35	0.00	0.00
	Anderson-Darling	0.00 / 0.00	0.00	0.42 / 0.27	0.02	0.01
	Lillefors	0.00 / 0.00	0.00	0.02 / 0.01	0.17	0.02
	Jacque-Bera	0.00 / 0.00	0.00	0.53 / 0.43	0.00	0.01
Log data	Shapiro-Wilk	0.00 / 0.13	0.35	0.00 / 0.12	0.80	0.81
	Anderson-Darling	0.00 / 0.10	0.44	0.01 / 0.17	0.49	0.66
	Lillefors	0.02 / 0.08	0.53	0.01 / 0.23	0.55	0.52
	Jacque-Bera	0.00 / 0.21	0.58	0.00 / 0.13	0.91	0.65

Significance levels assigned by four different normality tests to the GAPDH expression data. Low *P* values (<0.05) favor rejection of the normality hypothesis. In the case of the HSC and CMP/flk2$^+$ data sets, tests were conducted on the complete data set (first tabulated value), and on the data set with a single, very low outlier data point removed (second value). The tests were applied to both raw and log-transformed expression data. With log-transformed input, the assigned significance levels pertain to the hypothesis of lognormality rather than normality. The normality hypothesis is strongly disfavored for all except the CMP/flk2$^+$ data set. The CLP, CMP/flk2$^-$ and MEP data sets were highly compatible with a lognormal distribution. Lognormality scores reached significance for the remaining HSC and CMP/flk2$^+$ data sets if their lowest outlier data points were excluded from the analysis.

Supporting Text

Digital PCR Response Characteristic. If template molecules are randomly distributed within n compartments, the probability of any given molecule being trapped in any given compartment is $1/n$. Hence, the probability of a given compartment being empty when there are x template molecules in the sample is $(1 - 1/n)^x$. The expected number of non-empty compartments, y, is therefore as follows:

$$y = n[1 - (1 - 1/n)^x]$$

Consequently, a readout of y positive compartments gives the following estimate for x:

$$x = \log(1 - y/n) / \log(1 - 1/n)$$

If the number of positive reactions is small compared to the number of compartments, $x \gg y$. The response curve is therefore close to linear at low template concentrations. The statistical uncertainty in the estimated number of input molecules increases as the fraction of occupied compartments approaches unity. For a Digital Array panel with 1200 compartments, this error remains small even at an input of 4,000 molecules (CV » 5%, by Monte Carlo simulation). However, the shallowness of the response curve above 1,000 input molecules implies increased sensitivity to uncertainty in positive/negative calls.